Comparing Comparative Genomics in a Microbial Defensive Symbiont and Beyond

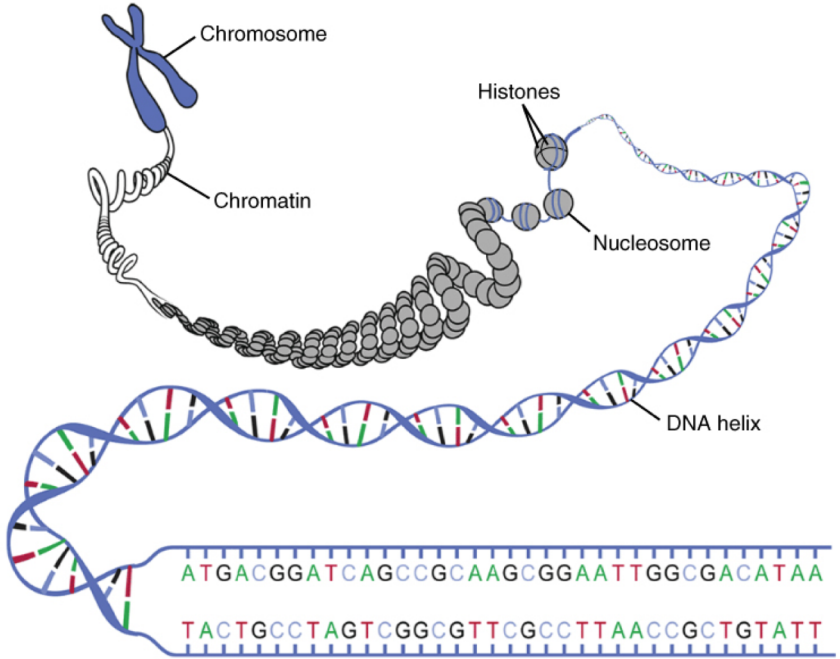
ComBEE 10/10/2019

Jenny Bratburd

Currie Lab

Comparative Genomics

- Genomic features (DNA, genes, synteny, regulatory sequences, structure)
- Tool to study evolution



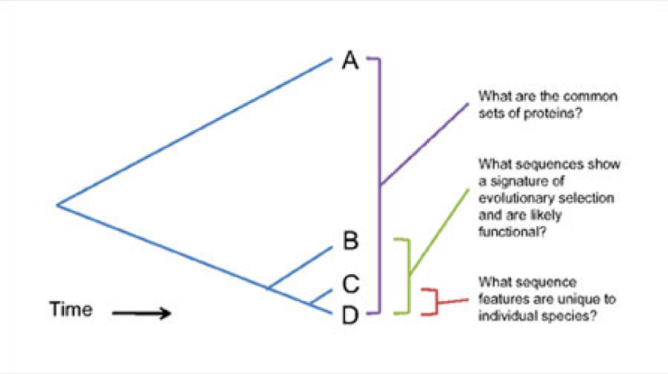
OpenStax Anatomy and Physiology

How to Compare Genomes

1. What scale should you compare?
2. Which genomes should you use?
3. What structural features and gene content can you compare?
4. Do you have phenotypes to compare?

What scale should you compare?

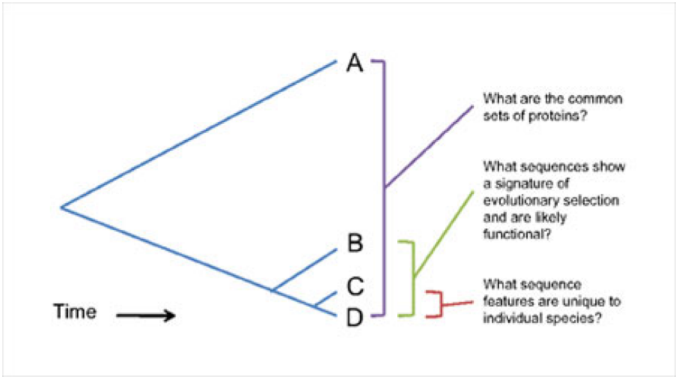
- Population level
  - Most specific
  - Unit of natural selection



Touchman 2010

What scale should you compare?

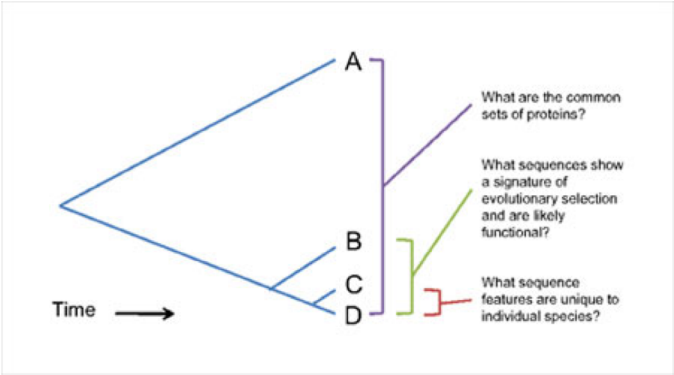
- Population level
  - Most specific
  - Unit of natural selection
- Species level
  - Smallest evolutionary independent lineage
  - Species concept



Touchman 2010

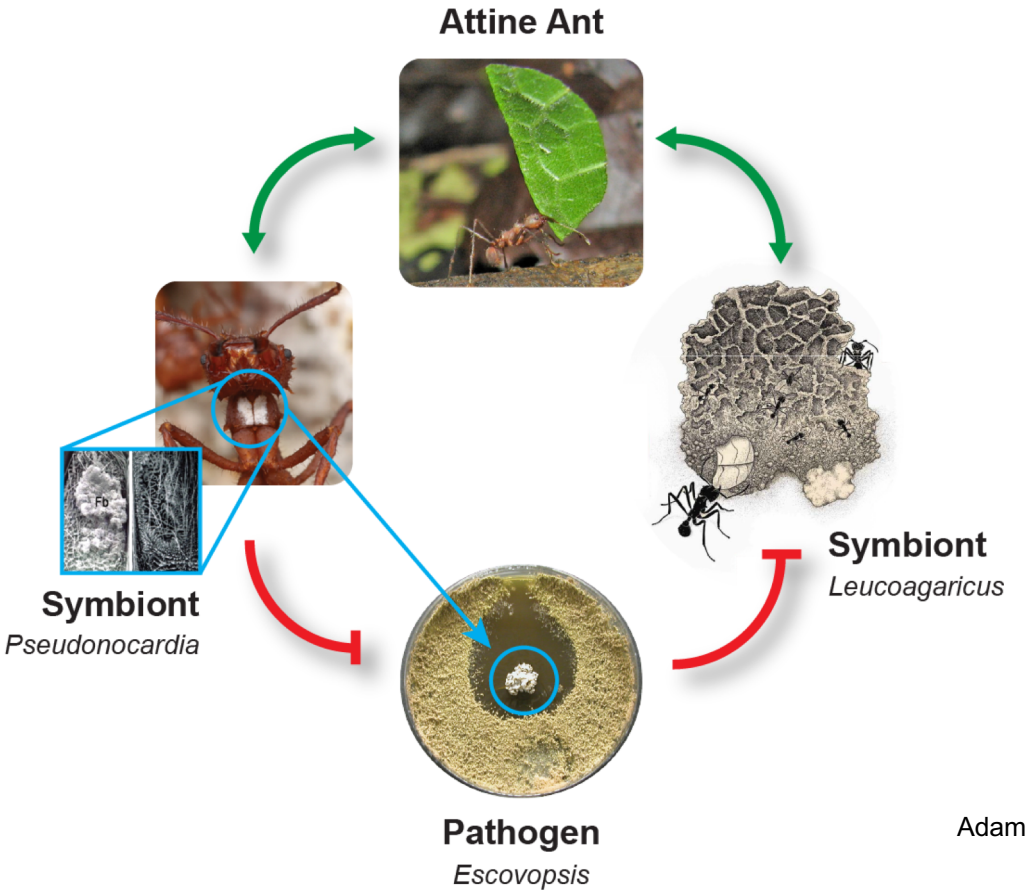
What scale should you compare?

- Population level
  - Most specific
  - Unit of natural selection
- Species level
  - Smallest evolutionary independent lineage
  - Species concept
- Genus, family, etc
  - More genomes available, but may have diverged millions of years ago
  - Taxonomic issues



Touchman 2010

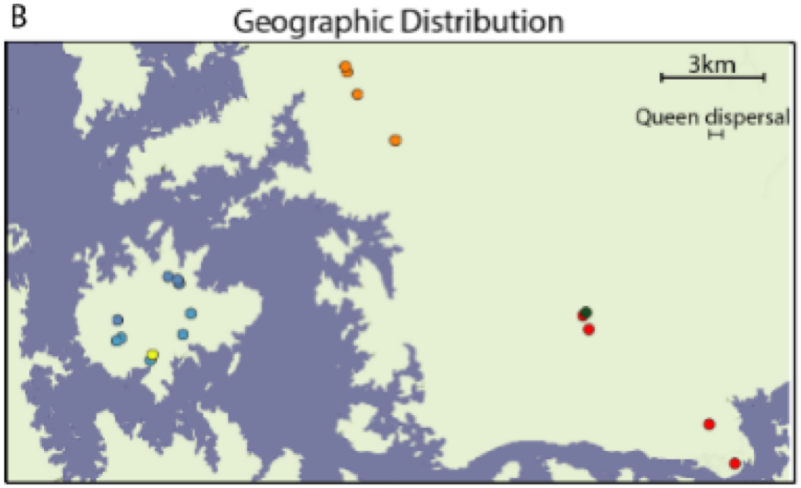
Leafcutter ants have a defensive mutualist

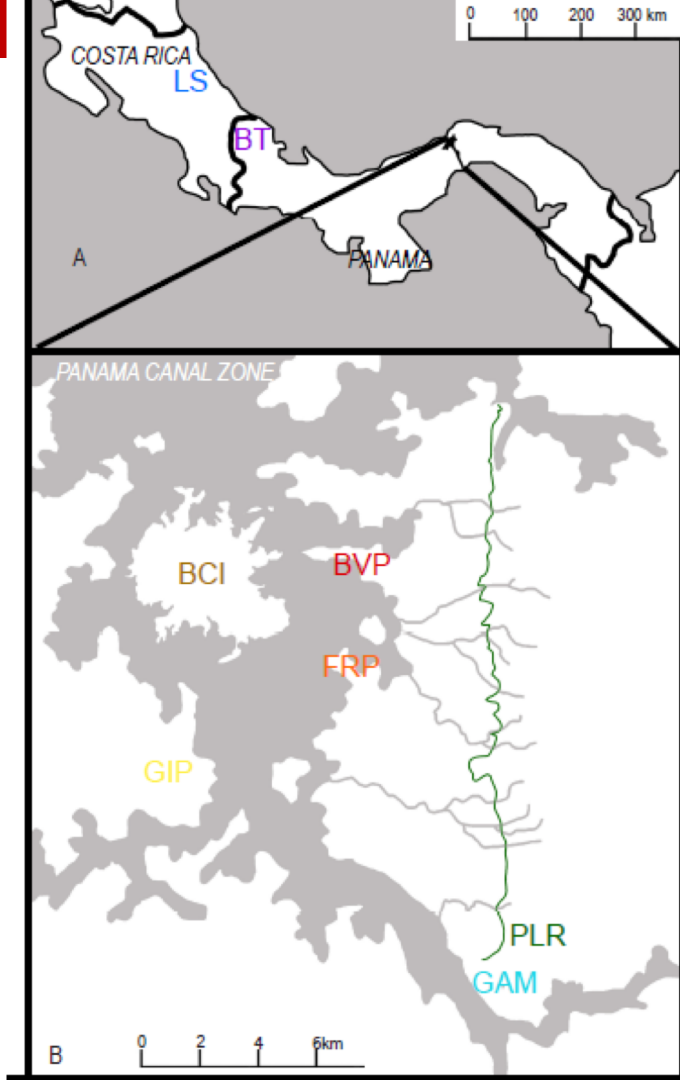


Adam Book

Population Scale Genomic Comparisons

Does genomic diversity vary over relatively small distances?

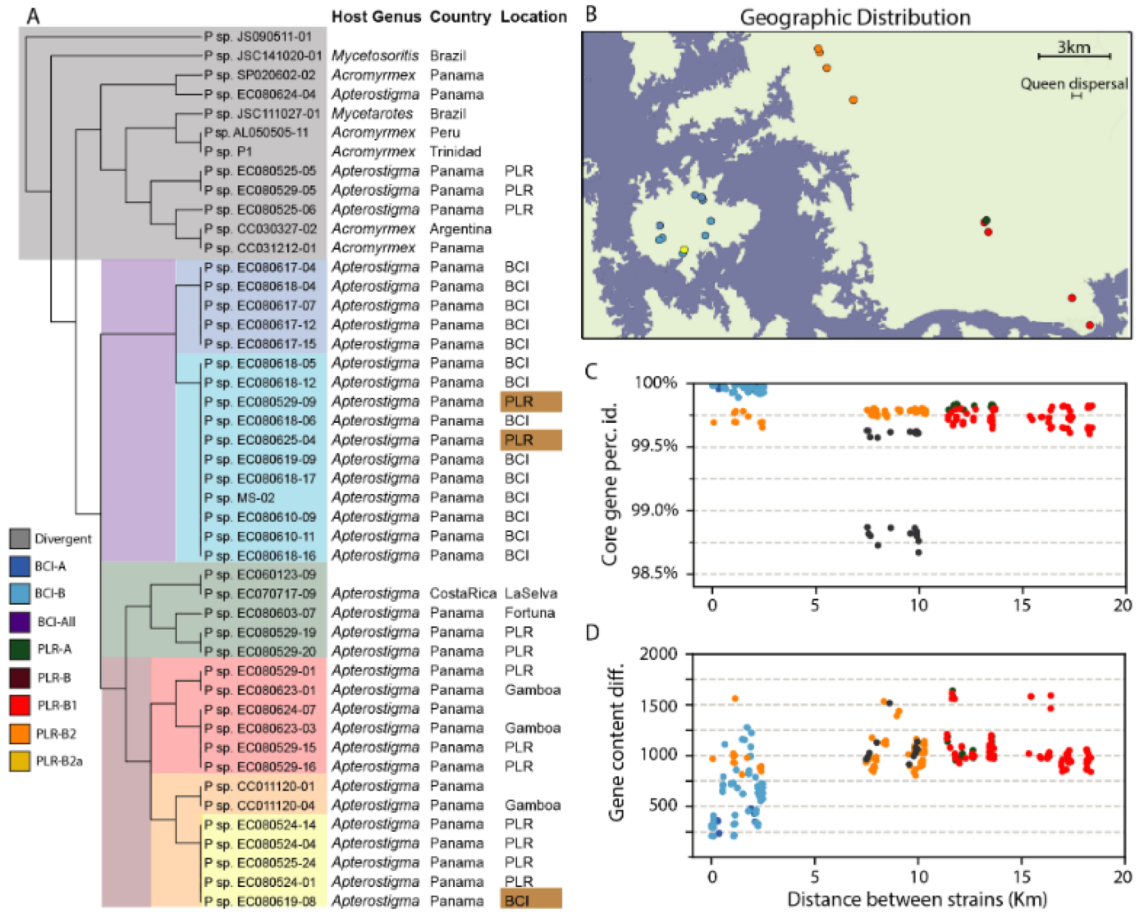




McDonald et al. bioarxiv https://doi.org/10.1101/545640

SNP identification with nucmer clustered with fineSTRUCTURE

Gene content diversity exists even within very closely related bacteria, can be structured by geography



McDonald et al. bioarxiv https://doi.org/10.1101/545640

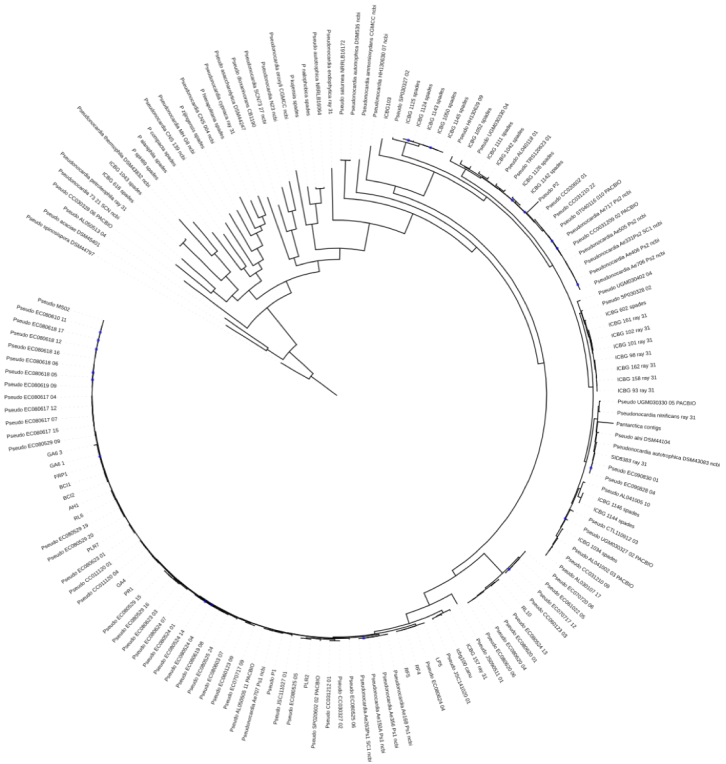
Genus level Comparison

Sources of Pseudonocardia including: industrial waste, soil, *Lobelia clavata* leaves, Acacia auriculiformis roots, and fungus-growing ants (Don Parsons)

- *Pseudonocardia* can be found in many sources, what gene content distinguishes ant-associated strains versus non-ant associated strains?

Trimming Genomes

Sources of Pseudonocardia including: industrial waste, soil, *Lobelia clavata* leaves, Acacia auriculiformis roots, and fungus-growing ants (Don Parsons)



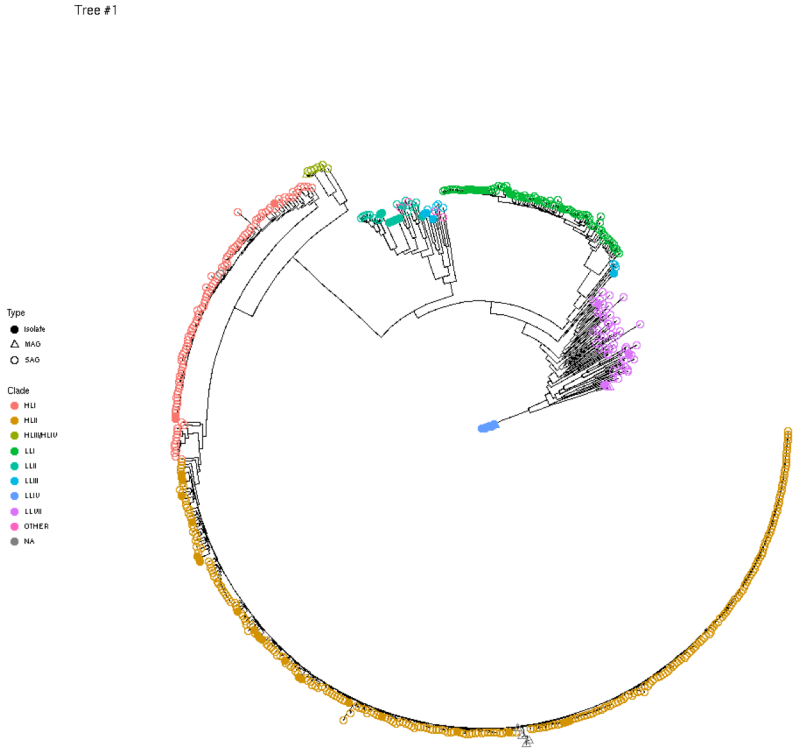
Which genomes should you use?

- Ideally group randomly and evenly sampled, with an abundance of high quality genomes

Which genomes should you use?

- Ideally group randomly and evenly sampled, with an abundance of high quality genomes
- Oversampled/undersampled strains of interest
  - Treemmer used to randomly prune tree while maintaining certain strains of interest

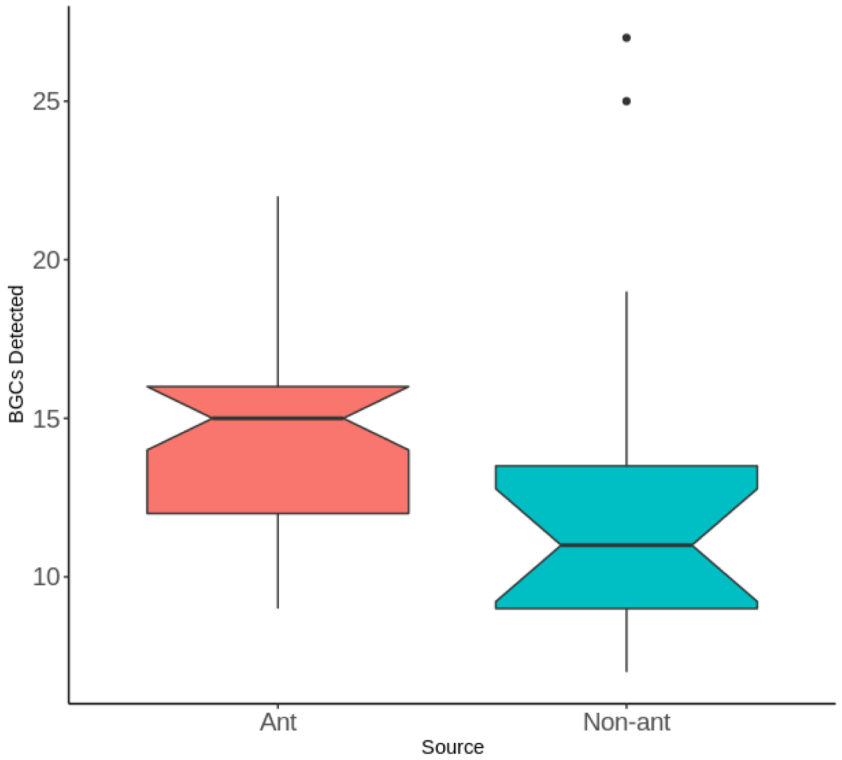
Trimming



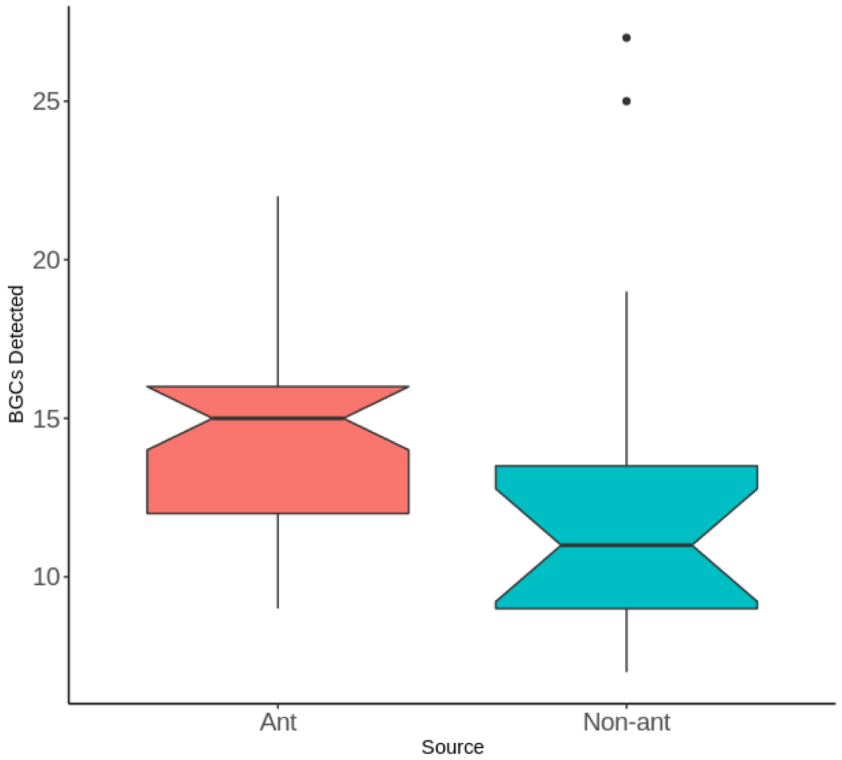
Treemmer + Treemmer animate
https://github.com/thackl/treemmer-animate

Which genomes should you use?

- Ideally group randomly and evenly sampled, with an abundance of high quality genomes
- Oversampled/undersampled strains of interest
  - Treemmer used to randomly prune tree while maintaining certain strains of interest
- Quality of genomes
  - Map to high quality reference
  - May change ability to address certain questions



Biosynthetic gene clusters detected with antiSMASH



Biosynthetic gene clusters detected with antiSMASH

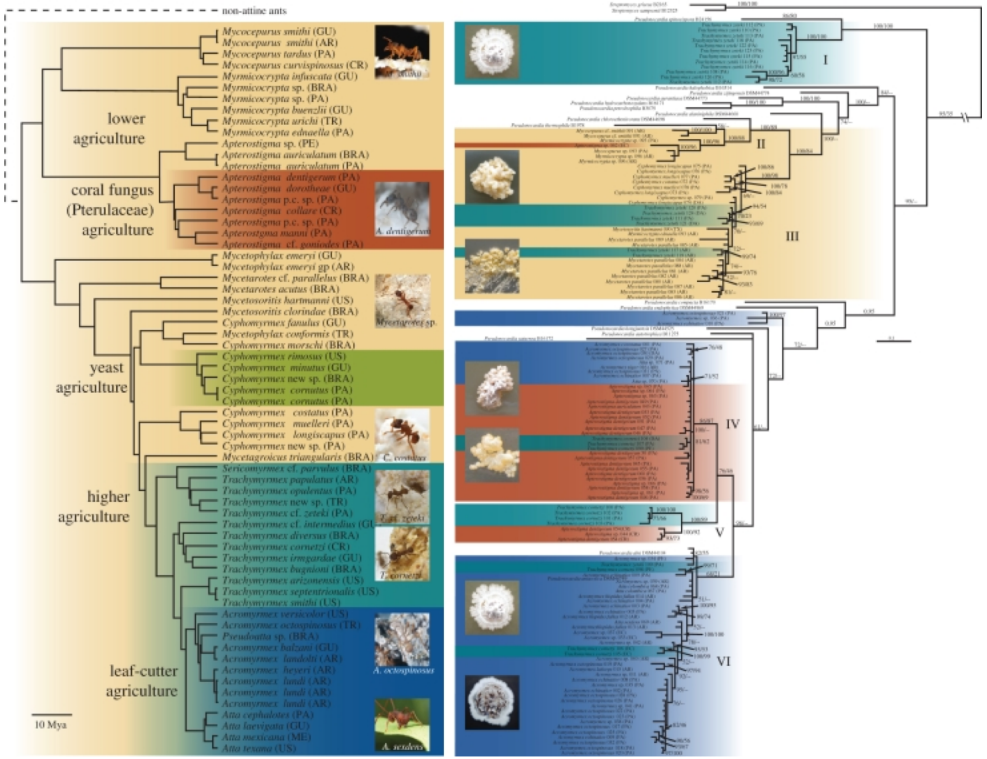
Not significant after removing low quality genomes

Which genomes should you use?

- Ideally group randomly and evenly sampled, with an abundance of high quality genomes
- Oversampled/undersampled strains of interest
  - Treemmer used to randomly prune tree while maintaining certain strains of interest
- Quality of genomes
  - Map to high quality reference
  - May change ability to address certain questions
- Quality of metadata
  - Metadata can be incomplete, misleading, or wrong
  - Availability of cultured strain for experiments

What can you compare?

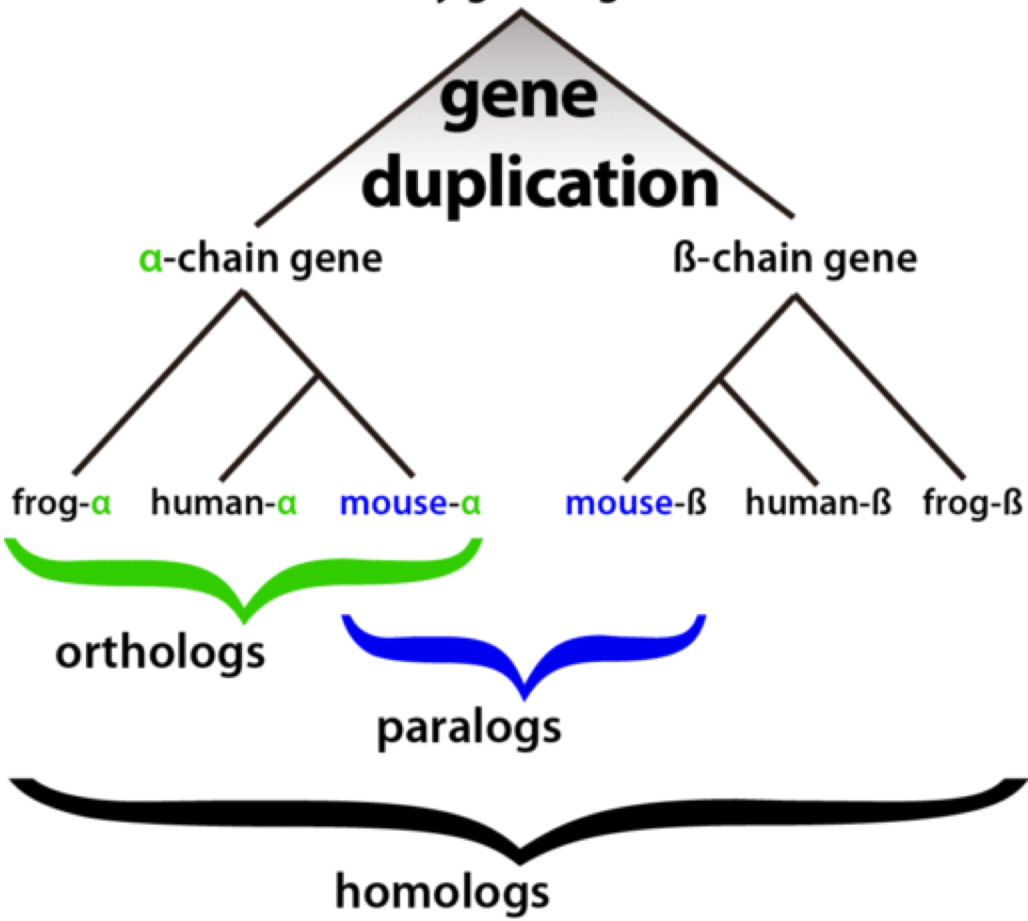
- Phylogeny
  - Multilocus, full genome
  - Coevolution



Cafaro 2010

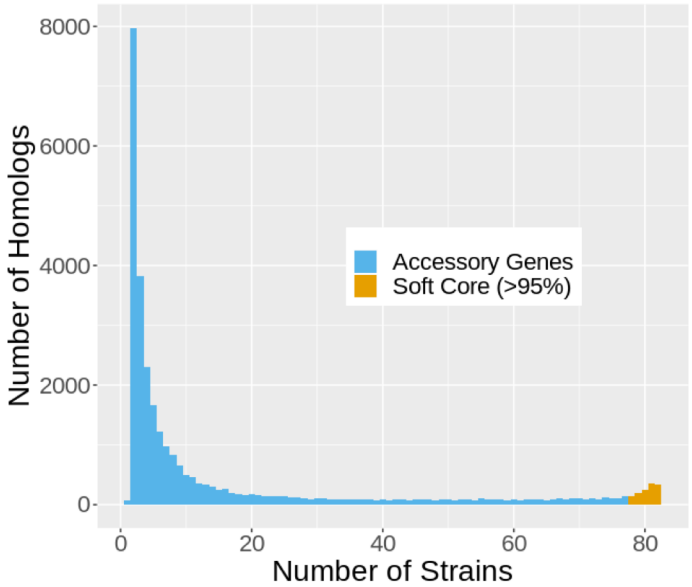
Homologs

- Homologous genes
  - OrthoMCL, proteinortho, pyparanoid, roary, gethomologues, anvio
  - Analyzing based on category KEGG/COG/etc
  - TreeWAS
  - Fisher's Exact Test



Popo H. Liao via Wikimedia Commons

Core Genome

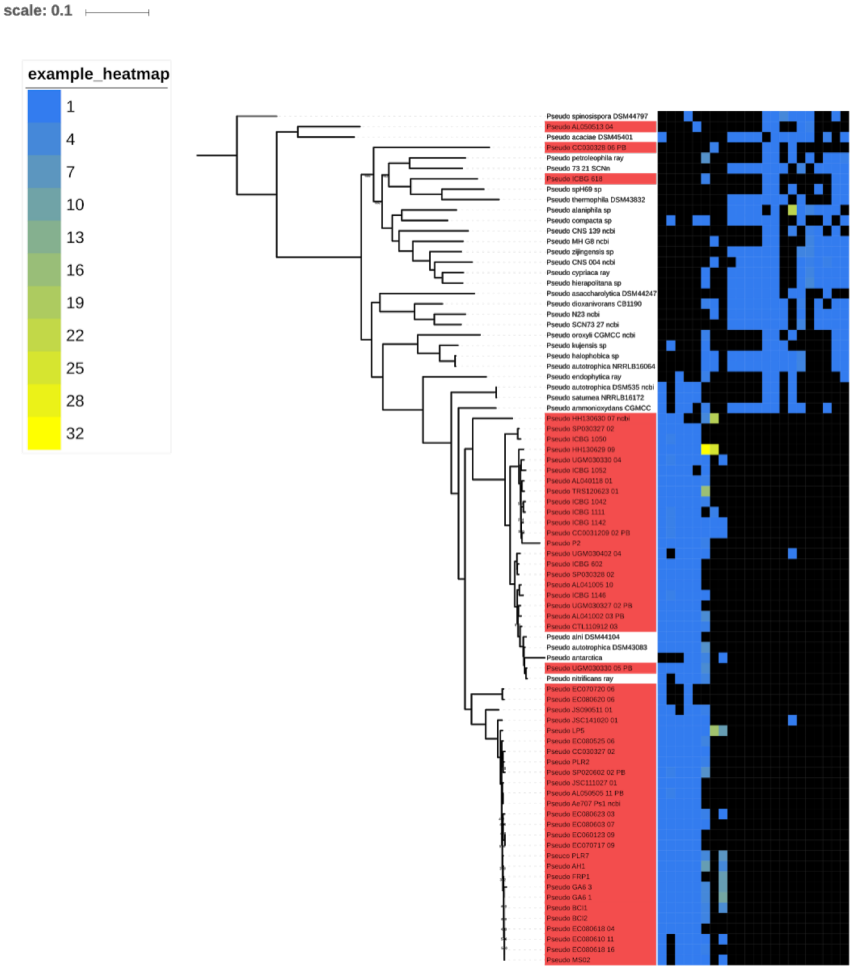


Pyparanoid

Diversity within *Pseudonocardia*

https://github.com/bratburd/comparative-genomics

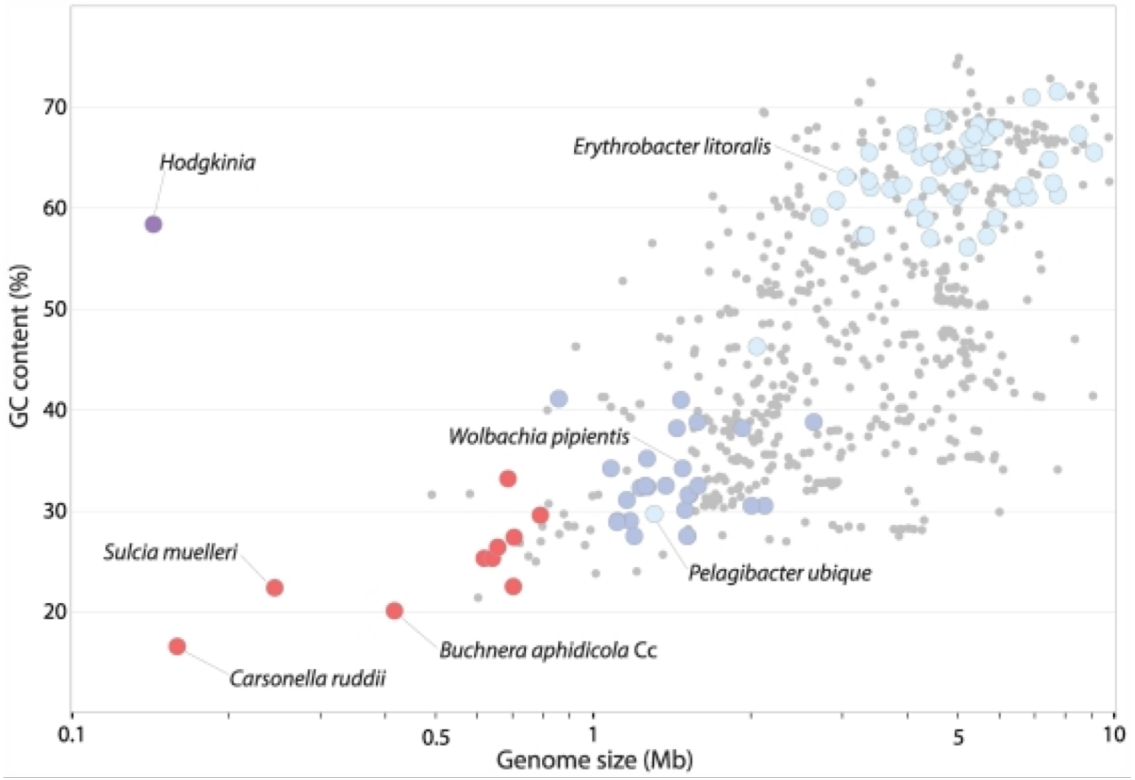
Example Enriched Genes



What else can you compare?

- Genome size, GC content

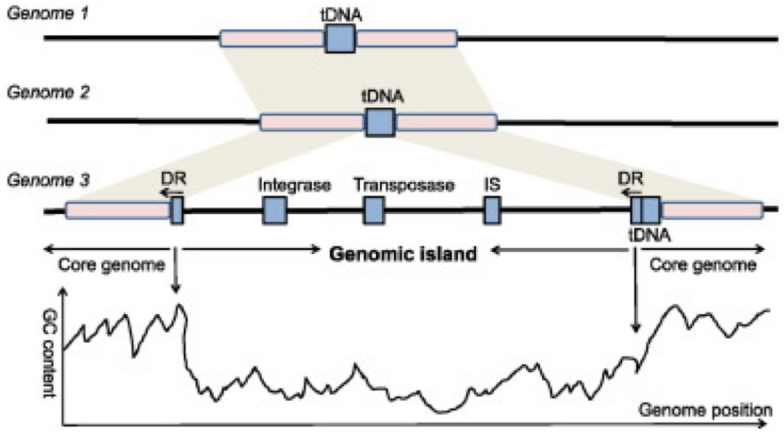
Comparing GC content and genome size revealed alternative stop codon



McCutcheon et al 2009

What else can you compare?

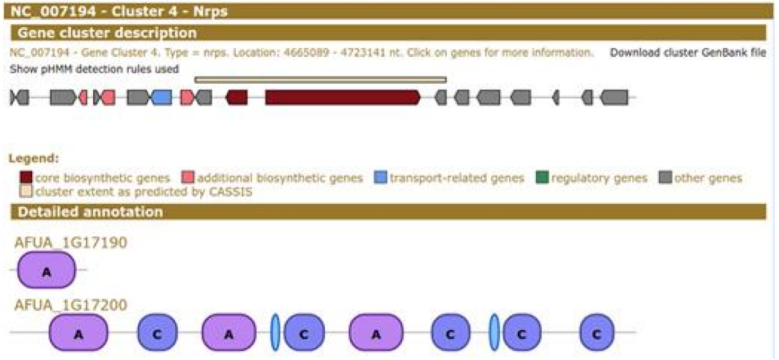
- Genome size, GC content
- Horizontal gene transfer
  - Plasmids (PlasmidFinder)
  - Phage (PHASTER)
  - Recombination (BratNextGen)
  - Genomic islands (Islander)



Lu and Leong 2016

What else can you compare?

- Genome size, GC content
- Horizontal gene transfer
  - Plasmids (PlasmidFinder)
  - Phage (PHASTER)
  - Recombination (BratNextGen)
  - Genomic islands (Islander)
- Biosynthetic gene clusters
  - antiSMASH, PRISM



Blin et al 2017

What else can you compare?

- Genome size, GC content
- Horizontal gene transfer
  - Plasmids (PlasmidFinder)
  - Phage (PHASTER)
  - Recombination (BratNextGen)
  - Genomic islands (Islander)
- Biosynthetic gene clusters
  - antiSMASH, PRISM
- Antibiotic resistance (ResFinder)
- Virulence factors (VFDB)
- Population genetics, positive selection

Do you have phenotypes to compare?

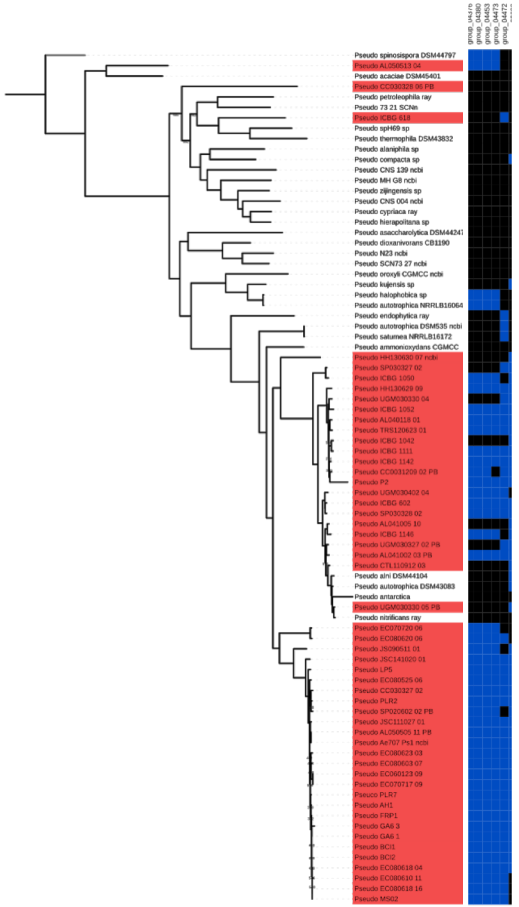
- Phenotypes are fun!
  - Antibiotic production/pathogen inhibition
  - Pathogenic versus mutualistic strains
  - Ability to colonize

- Non-ant associated *Pseudonocardia* can colonize ants
- Reanalysis may shed light onto genes needed for colonization

Differential Gene Content

Analysis based on source metadata

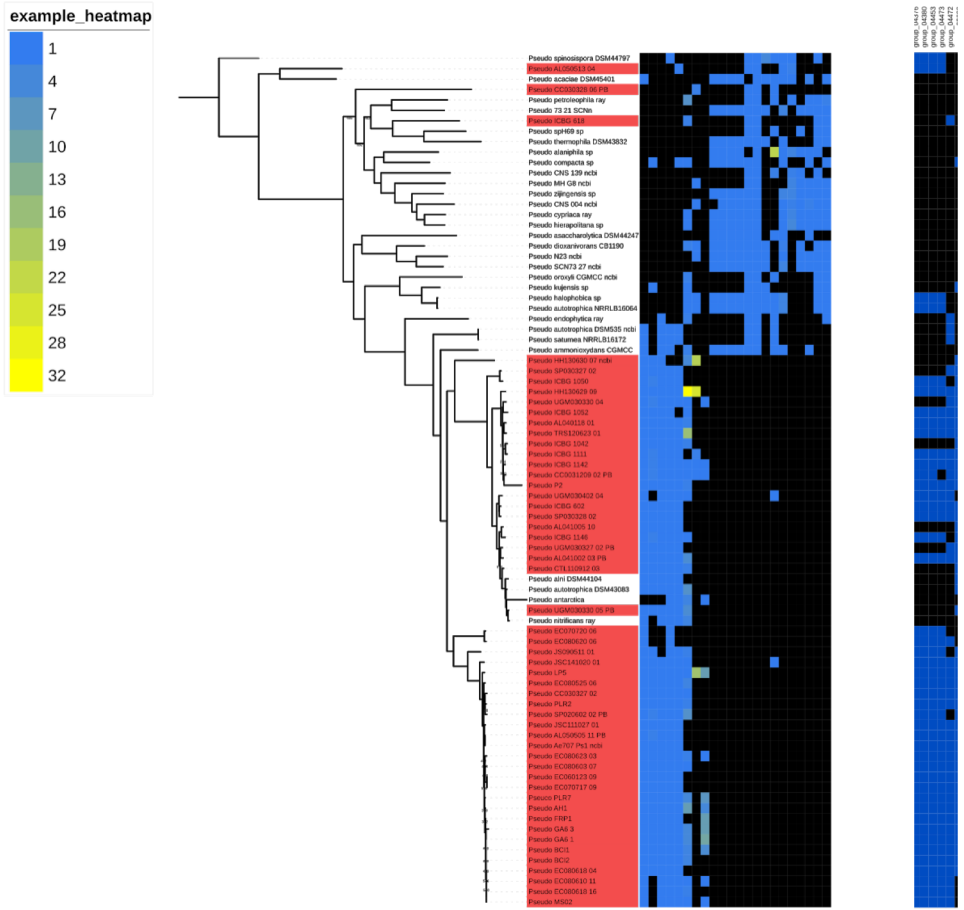
Pyparanoid
TreeWAS



Differential Gene Content

Analysis based on experimental metadata

Pyparanoid
TreeWAS



Which program should you use?

- Ease of installation and use
- Support and documentation
- Speed
- Accuracy

Different modes of 'reproducibility'

Methods	Same experimental system	Different experimental system
Same Methods	Reproducibility	Replicability
Different Methods	Robustness	Generalizability

Schloss, 2018 (10.1128/mBio.00525-18)

https://www.repro4everyone.org

Conclusions

- Identifying biological question can help define dataset and approaches
- Many new methods are being developed
- Explore!

Getty Images

ComBEE

Thursday October 17th: Edna Chiang, 16S Amplicon Sequencing Analyses